大数据时代的社会计算

沈华伟 程学旗

引言

随着智能终端、传感设备的普及和泛在数据感知技术、社交网络服务的快速发展，人类社会的信息化日益深入，积累下的大量社会感知数据为定量理解人类社会提供了前所未有的数据资源。遍布的监控摄像和传感设备实时获取 PB 级的数据，为我们认识物理世界的环境因素变化、城市交通状况等提供了详细的资料；Facebook 等社交网络和新浪微博等社会媒体每天记录着数亿用户的所言所行，为我们了解网络空间的舆情热点提供了丰富的数据；手机、手持电脑等移动智能终端记录着人类移动模式等实时社会动态，为我们理解人类的社会活动规律提供了原始素材。这些社会感知数据是连接物理世界、网络空间、人类社会的纽带。其数据规模以指数量级随时间增长，并呈现出多源头、跨媒体、复杂关联、持续变化等特点，如何有效利用社会感知数据进行社会计算，是大数据时代的一项重要研究课题。

对大数据时代社会计算的研究，核心问题主要包括三个方面：(1). 社会因素的可计算性：分析哪些社会因素可以通过社会感知数据进行度量和计算；(2). 社会行为的可预测性：分析自发、随机的个体行为背后的规律和模式，探索群体行为的涌现机理并分析其可预测性；(3). 社会计算的复杂性：集中体现在多种社会因素的耦合以及社会行为的幂率特征，使社会计算面临诸多不确定性因素。

本文围绕大数据时代的社会计算，从社会化推荐、影响力分析、网络结构分析、网络信息传播几个方面介绍我们在社会计算方面进行的一些研究工作，主要包括：社会化推荐的后验效用[1]、可扩展高精度的影响力最大化算法[2]、网络多类型结构规则分析[3]和微博消息流行度预测[4]四个方面。

社会化推荐中的后验效用

社会化推荐是指人和人之间的直接推荐，推荐的发送者和接收者均为真实的个体，而不是人类设计的推荐系统。在社会化推荐中，个体之间的社会影响力发挥着重要的作用。社会影响力是指一个人的意见如何通过社会关系影响到另一个人的行为。人与人之间相互影响的特性，在很大程度上决定整个网络的行为模式。探明其作用机制，对于理解社交网络上的交互行为、设计病毒式营销模型等是很有必要的。

我们的研究基于口碑推荐，一个人（发送者）向他的朋友（接收者）推荐某一商品，这会对接收者产生如何的影响。以往对这一问题的研究主要集中于“口碑推荐如何影响朋友的购买行为”，影响力体现在接收者做出决策前对该商品的先验期望的变化。但与之对应的另一方面目前很少有研究关注，即口碑推荐如何影响接收者做出决策后对该商品的后验体验的变化。比方说，当张三告诉李四某部电影值得一看，目前的研究关注这一推荐行为如何鼓励李四走进电影院买票观看，但甚少了解李四看完电影后的观感体验。通常的直觉认为，推荐行为不会对后验体验有影响，因为后验体验只取决于李四本人对电影的兴趣以及电影本身的水准。但本文发现了一个反直观的现象，推荐行为对于后验体验的影响是显著存在的。我们通过统计假设检验验证了这一现象，据此定义了后验的社会影响力并初步探索了其相关因素。

首先我们分析了在两个在线社交网络（豆瓣和 Goodreads）中，如果用户对某一对象（如电影/书籍/音乐等）做出评分之前曾经有朋友向他推荐过这一对象，他的评分是否会有显著改变。图 1(a)展示了一部示例电影的评分分布图，浅色柱表示接收过推荐的用户的评分概率，深色柱表示未接收过推荐的用户的评分概率。浅色柱在高分区域（5 分）的概率明显高于深色柱，表明用户在接收到推荐的情况下更倾向于对这部电影给出较高评分。图 1(b)和(c)分别展示了在豆瓣和 Goodreads 上对这一现象的统计结果。实线表示接收过推荐的用户的评分概率分布，虚线表示未接收过推荐的用户的评分概率分布，可以看到上述结果得到大样本统计支持。于是我们认为，“用户曾经接收到朋友对于某一对象的推荐”这一事实，与“用户给出较高的评分”，是具有明显的关联关系的，亦即“用户在体验过被推荐对象后的后验体验较好”。

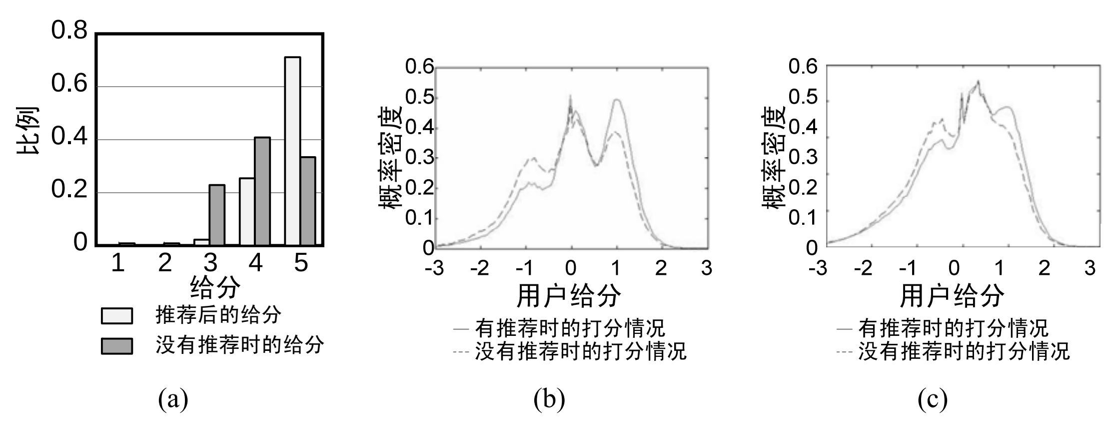


图1. 豆瓣/Goodreads 用户收到推荐和未收到推荐时对电影/书籍/音乐的打分情况

仅仅观测到关联关系是不够的，还需要进一步判断这种关联是否来自因果性，才能验证推荐行为是否能影响用户的后验体验。两个事件的相关性只可能来自两种情况：某一个事件是另一事件的原因；或者存在第三方因素作为这两个事件的公共原因。图 2(a)表示了一种可能的解释，“用户的后验体验”r 与“朋友是否曾给出推荐”m′是独立的两个事件，由于同时受到某个第三方因素（例如对象本身的质量、朋友之间的兴趣相似程度等）的影响而表现出关联关系，我们称其为“独立模型”。图 2(b)表示了另一种可能的解释，r 直接受到 m′的影响，我们称其为“影响模型”。

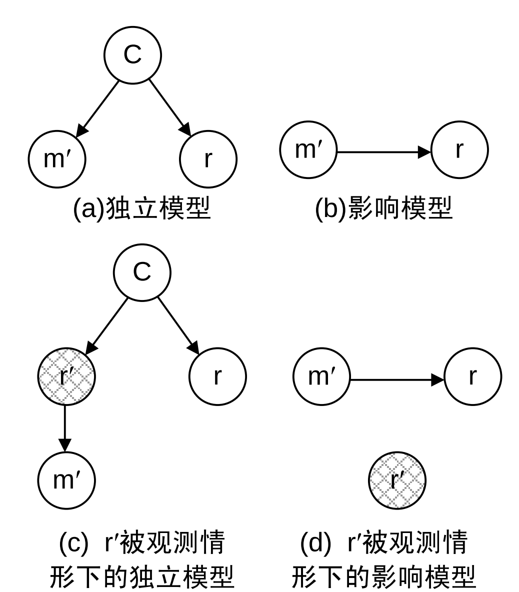


图2. 关联关系的两种可能解释

为了说明哪一种解释是合理的，我们引入另一个观测量——推荐人自己的给分 r′（如果没有推荐人，即 m′=0，则随机挑选一个用户的给分作为 r′） 来简化模型。不妨假设 m′完全由 r′决定，即每个人是否做出推荐完全取决于他自己对该对象的后验体验。独立模型变成图 2(c)所示情况。这个模型存在一种条件独立性关系，即当 r′被观测的时候，m′与 r 是条件独立的。于是影响模型变成了图 2(d)所示情况。由于 r′不参与影响模型，因此 r′被观测的时候，m′与 r 仍然是相关的。据此我们只要观察在真实数据中，r′被观测的情况下 m′与 r 的条件独立性关系即可判定哪种解释是合理的。

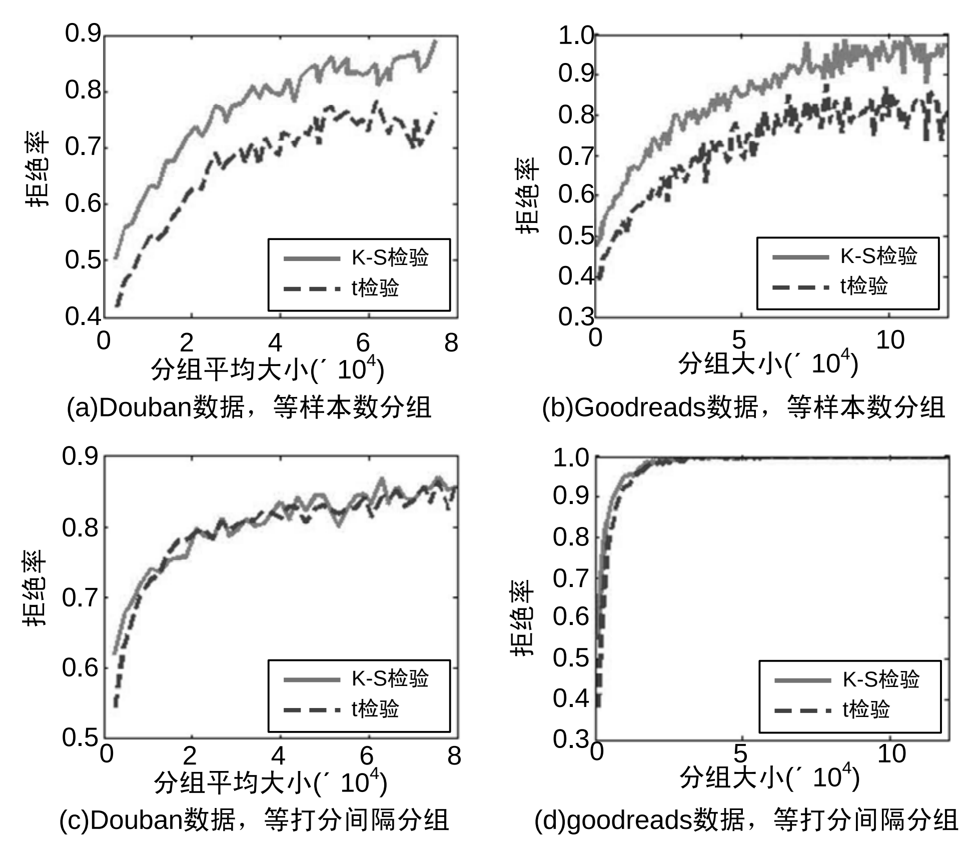


图3. 统计假设检验结果

我们设计了统计假设检验来验证上述条件独立性关系。如图 3 所示，在两个数据集和两种不同的离散化处理方式下，检验结果表现一致，均以较高概率拒绝同分布零假设，即认为独立模型中的条件独立性关系不成立。因此我们得出结论，“朋友是否做出推荐”会直接影响“用户对于被推荐对象的后验体验”。我们据此定义用户的社会影响力为“做出推荐后，接收推荐的用户的后验体验的改变程度”。两种用户常常被发现具有这样的影响力，一种用户拥有很多的朋友，他们的话语通常因具有较高的权威性而被接受，另一种用户拥有敏锐的鉴赏能力，能够先于大多数普通用户发现优质对象。

静态影响力最大化算法

影响力最大化（Influence Maximization, IM）问题是在给定传播模型的基础上，解决如何在网络上选择一部分初始用户，由他们通过口口相传的口碑效应将企业的产品或信息尽可能地推广出去。肯培（Kempe）等人最早将将该问题形式化为一个离散型优化问题[5]：给定一张由社会网络抽象出的图，一个影响力传播模型，和一个整数 $\boldsymbol{k}$，要求在图中寻找一个由 $\boldsymbol{k}$ 个节点（也称为种子节点）组成的集合 $\boldsymbol{S}$，使得该集合 $\boldsymbol{S}$ 在当前影响力传播模型下，期望的影响力传播规模（即最终被影响成功的节点总数）尽可能最大。他们证明了独立级联模型和线性阈值模型上 IM 问题的目标函数具有单调性和子模性，因此使用贪婪算法进行求解能取得一个较优的近似解，近似比为 1-1/e（约为 0.63）。贪婪算法需要我们能够计算出给定节点集合的影响力，然而精确计算出给定节点集合的影响力具有很高的计算开销，因此通常采用蒙特卡罗模拟的方式进行近似计算。

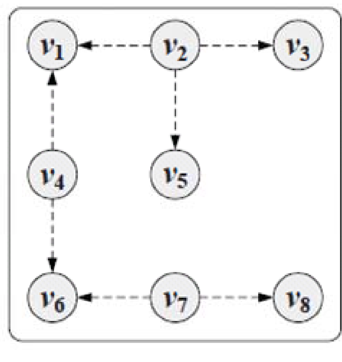
为了提高贪婪算法的可扩展性，研究人员提出了一系列的改进策略，包括 CELF[6]、CELF++[7]和 NewGreedy[8]。CELF 利用了 IM 问题目标函数的子模性，从第二轮起每轮只需要检查少量的侯选节点，从而有效地降低了计算量。CELF++又在前者基础上充分利用单轮

蒙特卡罗模拟去同时计算两个集合的影响力，在仅增加少量内存的前提下有效地减少计算量。NewGreedy 通过对网络中所有边进行预判定的方式达到在每轮蒙特卡罗模拟中同时检查各候选节点性能的目的，但这种方法相比 CELF 仅在第一轮计算中具有优势，从而作者又提出了 MixedGreedy，在第一轮中采用 NewGreedy 的方法，第二轮往后均使用 CELF。上述这些方法都有效地降低了自然贪婪算法的计算复杂度，同时又基本保证了解的精度，但依旧无法适用于动辄上百万节点、上亿条边的社交网络。为了进一步提高求解 IM 问题的速度，人们开始致力于设计扩展性强、高效的启发式算法[9]。但是启发式算法不能像贪婪算法那样保证问题的求解精度，结果的可靠性没有保证。

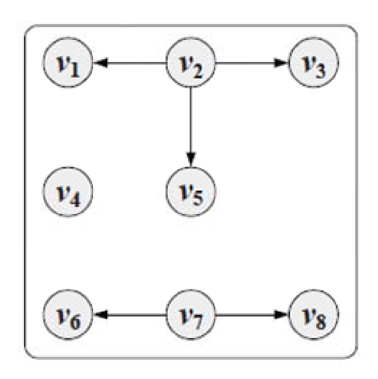
我们从贪婪算法入手研究高精度可扩展的 IM 问题算法。我们指出，贪婪算法的可扩展性和高精度无法兼顾的原因在于：为了保证贪婪算法的精度，需要尽可能准确地计算出给定节点集合的影响力，这需要进行很多次数的蒙特卡罗模拟，从而导致算法的可扩展性差。

针对贪婪算法面临的“高精度”与“可扩展”的矛盾，我们分析发现：现有的贪婪算法由于采用蒙特卡罗模拟来近似计算给定节点集合的影响力，结果导致，在贪婪算法的计算过程中，IM 问题的目标函数不再具备模块性和单调性的特征。为了克服该问题，现有的贪婪算法大多采用提高蒙特卡罗模拟的次数来尽可能保证 IM 问题目标函数在贪婪算法过程中的模块性和单调性，一旦降低蒙特卡罗模拟的次数，模块性和单调性就无法保证，导致算法的精度降低。

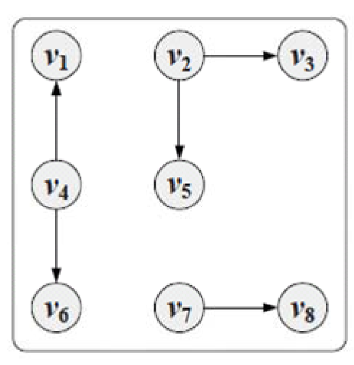
在发现“高精度”与“可扩展”矛盾的症结之后，针对独立级联传播模型，我们提出了一种静态贪婪算法，在求解 IM 问题时可以兼顾算法的高精度和可扩展性。具体而言，我们利用独立级联模型的性质，对每条边上的传播概率进行独立采样，从而得到一个传播网络。计算给定节点集合的影响力，等价于在传播网络上找给定节点的可达节点范围。最后，通过多次独立采样，得到多个传播网络，将各个网络上计算出的影响力求均值，作为给定节点集合的影响力。如此一来，在贪婪算法的计算过程中，这些事先得到的传播网络被重复用来计算各个节点结合的影响力，从而严格保证了 IM 问题目标函数的模块性和单调性。在模块性和单调性得到严格保证的情况下，我们只需要少数几次独立采样，即可充分估计出节点结合的影响子，通常只需要 100 次左右蒙特卡罗模拟，相比于现有贪婪算法的 20000 次而言，降低了 2 个数量级。



(a) 影响力网络



(b)传播网络 1



(c)传播网络 2

图4. 现有贪婪算法不能保证模块性的示意图

图 4 通过示意解释了现有贪婪算法不能保证模块性的原因。子图(a)中是独立级联模型对应的影响力网络，虚线表示这条边以某个概率存在。子图(b)和(c)是对子图(a)中的网络进行采样得到的传播网络。其中，子图(b)中的传播网络在贪婪算法的第一轮用于计算节点集合的影响力，子图(c)中的传播网络在贪婪算法的第二轮用于计算节点集合的影响力。贪婪算法在第一轮中会选定影响力最大的节点 v2 作为种子节点。注意，在第一轮中， v4 的影

响力为 0，因此 v4 的影响力边际效应为 0。在第二轮中，由于第一轮已经选择了节点 v2 作为种子节点，将节点 v4 作为新增减的种子节点，所带来的影响力边际效应为 1，即影响到了节点 v6。因此，节点 v4 在第二轮的边际效应比第一轮要大，这与模块性所要求的边际效应递减是矛盾的。因此，现有贪婪算法不能保证模块性。

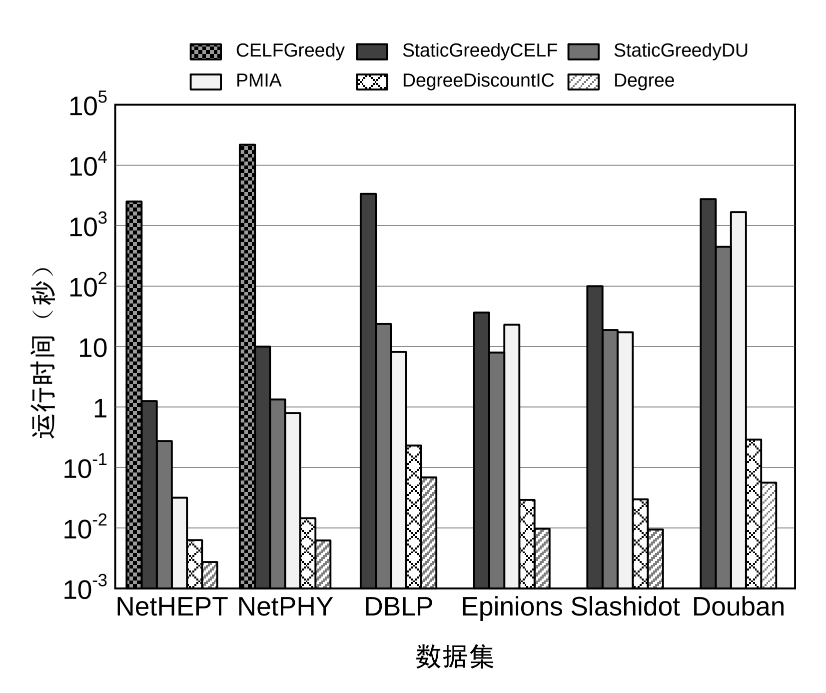


图5.　算法时间开销对比

我们在三个科学家合作网络（NetHEPT、NetPHY、DBLP）和三个社交网络（Epinions、Slashdot、Douban）上进行了实验。图 5 为算法的时间开销对比。从实验结果可以看出，我们提出的算法（StaticGreedy）大大降低了时间开销，和目前最快的启发式算法（PMIA）一致。另外，我们的算法采用了贪婪算法的框架，这在理论上保证了其计算精度和精度最好的贪婪算法（CELFGreedy）是一致的。综上所述，我们提出的静态贪婪算法通过严格保证 IM 问题目标函数的模块性和单调性，巧妙解决了现有贪婪算法面临的“高精度”和“可扩展”之间的矛盾。

网络多类型结构规则分析

探索网络结构和功能之间的关系一直是网络分析的重要研究内容。在过去十年间，社区结构作为很多真实网络所普遍具有的一种结构规则，得到了很多领域学者的广泛关注和深入研究。然而，除社区结构外，网络还具有多种类型的结构规则，包括多部结构（multi-partite structure）、层次结构（hierarchical structure）和核心-外围结构（core-periphery structure）等。这些多种类型的结构从不同侧面反映了网络的结构规则，而分析网络固有的结构规则对于我们认识网络和有效利用网络具有重要理论意义和实际价值。

我们针对网络多类型结构规则的发现展开研究，指出现有方法存在的两个不足之处：(1). 现有方法大多“先入为主”地假定网络具有某种特定类型的结构规则，基于这样的假定，进而设计算法去揭示该结构规则。因此，当对网络结构规则的事先假定与实际不符合时，算法往往无法正确地揭示网络的结构规则，甚至会得出错误结论；(2). 很多真实的网络往往同时具有多种类型的结构规则，而现有的方法却大多仅能揭示网络某种特定的结构规则。另外，网络有可能具有一些未知类型的结构规则，一个好的方法应该能够揭示出未知的结构规则。

现有方法存在的不足主要源于其对结构规则的定义缺乏灵活性，描述能力不足。例如，针对社区结构而设计的方法，认为社区是内部节点之间连接紧密、社区之间连接稀疏的节点结合，这样的定义局限于同配结构（assortative structure），因此无法适用于发现诸如多部结构在内的异配结构（disassortative structure）。

针对现有方法的不足，我们提出一种网络多类型结构规则的探索性分析方法。该方法把网络结构规则定义为"网络节点可以分成一些节点组，同一组内的节点具有相近的连接偏好或连接模式"。这一灵活的定义使得多种类型的结构规则可以在一个统一的框架下得以揭示。进而，我们把网络结构规则视为未观测到的量（hidden quantity），基于观测到的网络数据（节点间的连边）和对网络结构规则的定义，通过统计推断的期望-最大化（Expectation Maximization Algorithm）算法，推断出网络固有的结构规则。和现有方法相比，我们方法最大的优势在于其灵活性，这种灵活性使得我们的方法在克服现有方法不足的同时可以吸收他们的优点，揭示网络多种类型的结构规则。

具体地讲，我们的方法是一种随机分块模型（stochastic block model）。我们假定网络的 n 个节点被划分在 c 个模块内，随机选择的一条边 e_{ij}（表示由节点 i 指向节点 j 的边），其连接模块 r 和模块 s 的概率由 ω_{rs} 表示。另外，一条从模块 r 指出的边，其尾节点是节点 i 的概率由 θ_{ri} 表示，一条指向模块 s 的边，其头结点是节点 j 的概率由 φ_{sj} 表示。在我们的模型中，所使用的量可以分为三类：(1). 已观测到的量 A（网络邻接矩阵，其元素 A_{ij} 表示由节点 i 指向节点 j 的边的权重）；(2). 未观测到的量 $\vec{g}_{ij}$（表示边 e_{ij} 的尾节点 i 所来自的模块）和 $\bar{g}_{ij}$（表示边 e_{ij} 的头节点 j 所来自的模块）；和(3). 模型参数 ω_{rs}、θ_{ri} 和 φ_{sj}。根据我们的模型，一条边 e_{ij} 的生成的过程可以描述如下：

1. 以概率 ω_{rs} 选择两个模块 $\vec{g}_{ij}=r$ 和 $\bar{g}_{ij}=s$；
2. 从模块 r 中，以概率 θ_{ri} 选择节点 i 作为边 e_{ij} 的尾节点；
3. 从模块 s 中，以概率 φ_{sj} 选择节点 j 作为边 e_{ij} 的尾节点。

遍取未观测量 $\vec{g}_{ij}$ 和 $\bar{g}_{ij}$ 的所有可能值，观测到边 e_{ij} 的概率表示如下：

$$\mathrm{Prob}\left(e_{ij}\middle|\omega,\vartheta,\phi\right)=\sum_{rs}\omega_{rs}\theta_{ri}\varphi_{js}$$

进而，根据我们的模型，观测到整个网络的概率为：

$$\mathrm{Prob}\left(A\middle|\omega,\theta,\varphi\right)=\prod_{ij}\left(\sum_{rs}\omega_{rs}\theta_{ri}\varphi_{js}\right)^{A_{ij}}$$

另外，模型参数满足如下约束条件：

$$\sum_{r=1}^{c}\sum_{s=1}^{c}\omega_{rs}=1,\quad \sum_{i=1}^{n}\theta_{ri}=1,\quad \sum_{j=1}^{n}\varphi_{sj}=1$$

使用期望最大化算法（EM 算法），通过使根据模型观测到网络的概率最大化，我们得到

$$q_{ijrs}=\frac{\omega_{rs}\theta_{ri}\varphi_{sj}}{\sum_{rs}\omega_{rs}\theta_{ri}\varphi_{sj}}$$

和

$$\omega_{rs}=\frac{\sum_{ij}A_{ij}q_{ijrs}}{\sum_{ijrs}A_{ij}q_{ijrs}};\quad \theta_{ri}=\frac{\sum_{js}A_{ij}q_{ijrs}}{\sum_{ijs}A_{ij}q_{ijrs}};\quad \varphi_{sj}=\frac{\sum_{ir}A_{ij}q_{ijrs}}{\sum_{ijr}A_{ij}q_{ijrs}}$$

这里，隐变量 q_{ijrs} 表示观测到的一条边 e_{ij}，其尾节点 i 来自模块 r，结点 j 来自模块 s 的概率。

上面的两个式子构成了 EM 算法的核心，迭代计算上面两个式子直到收敛，便得到了模型参数 ω_{rs}、θ_{ri}、φ_{sj}、以及隐变量 q_{ijrs} 的具体取值。这些值提供了网络结构规则的所有信息。

图 7 对比了我们的模型和现有的两个代表性模型。其中，纽曼（Newman）的模型[10]可以发现社区结构和多部结构等多种类型的结构规则，然而由于缺少模块间关系的描述，使得其不能对所揭示的结构规则的类型进行有效判定，而且当网络中多种结构规则并存时无法有效识别网络结构规则。任（Ren）的模型[11]是我们模型的特例，仅能发现网络的社区结构，而我们的模型可以有效发现网络中多种类型的结构规则。

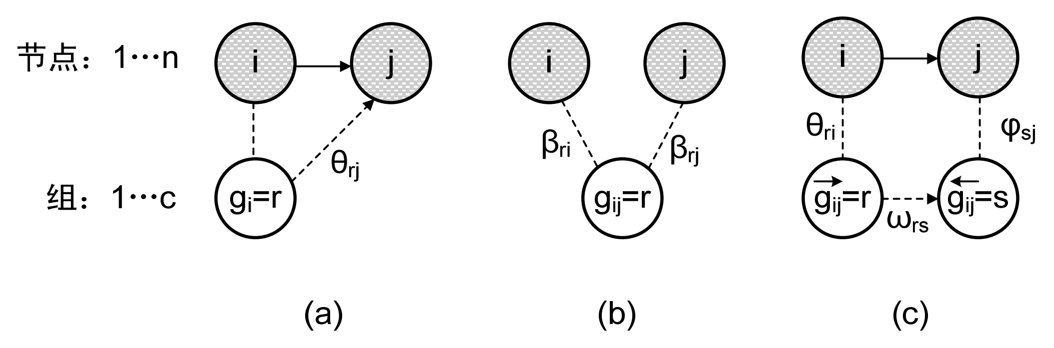


图6. 模型比较：(a) 纽曼的模型；(b)任的模型；(c)我们的模型

微博消息流行度预测

近年来，在线社交网络服务已逐渐成为信息网络应用的主流。类似 Facebook（脸书）、YouTube（优兔）、Twitter（推特）、新浪微博、人人网等社交网络不仅为人们提供了交互的社交平台，也在塑造现今互联网产业的商业模式上发挥了越来越重要的作用。涌现的这些社交网络带给人们分享信息和交互便利的同时，也给信息的过滤和有效利用提出了挑战。

随着大数据时代的到来，社交网络上海量的用户行为轨迹信息可以被获得和利用，这也为我们深入研究社交网络上信息扩散相关的一系列科学问题提供了机会。我们以新浪微博为例，将用户在新浪微博上发布的微博称为消息，并将微博被转发的次数称为消息的流行度。具体研究问题为：如何根据消息被发布后一个小时的扩散情况，预测其未来可能的流行度。该研究具有着重要的技术、商业和社会意义与价值：(1). 从技术的角度看，对消息流行度演化的理解，可以驱动服务提供商设计出具有成本效益的缓存和内容分发机制系统，以及发现诸如搜索引擎等系统中的潜在瓶颈；(2). 从商业的角度看，对消息流行度的预测不仅可以帮助新闻记者、内容提供商、广告商、新闻推荐系统等提供信息服务和病毒式营销策略，还可能辅助发现线上或线下的潜在商业机会；(3). 从社会的角度看，对消息流行度预测的深入研究，可以揭示人类群体行为的属性和规则，便于管理者更加及时准确地掌握、监管和引导公共舆论。

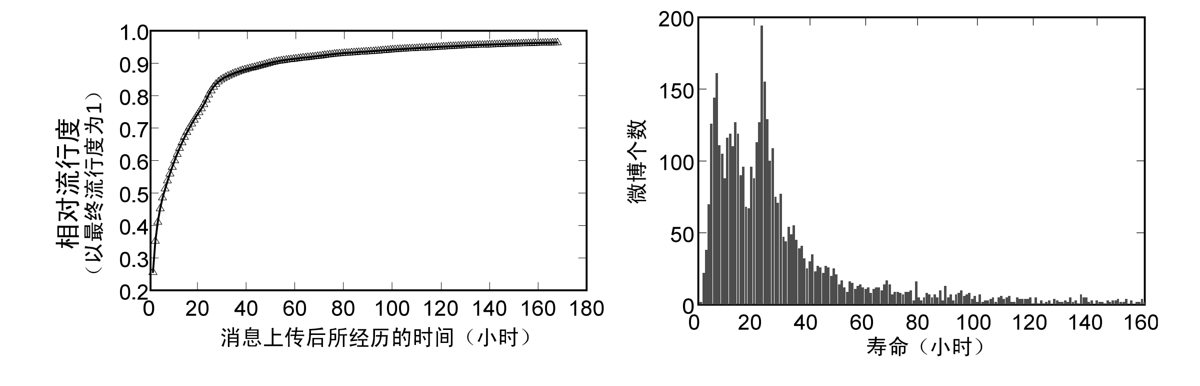
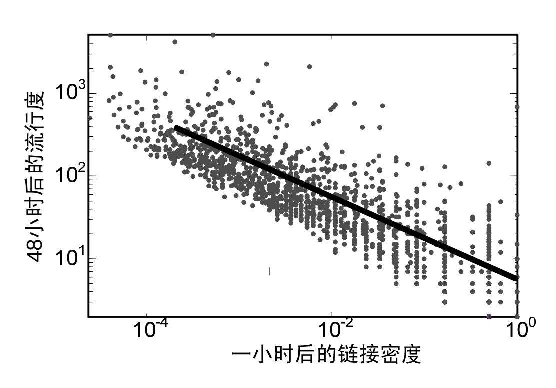


图7. 消息流行度和消息生命周期

我们首先对全部消息的流行度进行统计，发现消息流行度也是近似符合幂律分布的。这

表明少数消息获得了人们大量的关注，而大多数消息获得的关注都非常有限。这也是符合现阶段研究中人们对人类行为规律的认识的。通过对消息生命周期的分析，我们发现大多数消息的流行度在 24 小时内便达到了最终流行度的 80%，并在 48 小时内达到 90%，如图 7 左边部分所示。基于这样的统计发现，我们将消息的生命周期定义为 48 小时，并发现消息生命周期近似服从对数正态分布（log-normal distribution），如图 7 右边部分所示。

消息流行度预测的经典方法是利用 Digg[1]上消息早期与晚期的流行度的对数之间的强关联性，采用线性回归模型直接外推进行预测[12]。而这种强关联性在新浪微博上是否存在尚且未知，于是我们进一步分析了新浪微博上消息流行度的时序关联性。实验结果表明，新浪微博上消息早期与晚期的流行度的对数之间的皮尔森相关系数为 0.74，远低于经典方法中的接近 0.9。这个结果表明，针对新浪微博上消息流行度预测问题，经典的直接外推的方法未必适用。



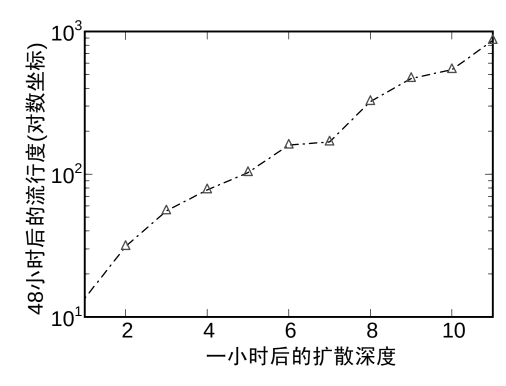


图8. 消息最终流行度和早期连边密度和扩散深度之间的关系

表 1 三种算法结果对比

方法	均方根差	绝对平均误差
基准方法	0.77	0.57
利用链接密度	0.63	0.45
利用扩散深度	**0.61**	**0.43**

本文针对信息扩散早期的传播者，通过挖掘他们之间的结构属性，试图为消息流行度的预测提供一些指示因素。直观上理解，如果一条消息早期的传播者之间在结构上具有多样性，那么它最终会扩散到更广范围的可能性就越大。本课题从消息早期扩散深度和传播者间连边密度两个角度来刻画结构多样性。通过实证发现，消息最终流行度的对数与早期传播者间连边密度的对数之间存在很强的负相关，与消息早期扩散深度之间存在很强的正相关，如图 8 所示。

基于上述实证发现，我们提出了融合结构属性的线性模型预测方法，采用均方根误差和平均绝对误差两种评估策略，与经典基准方法进行预测性能的比较，实验结果如表 1 所示。

综上所述，本课题针对社交网络上消息流行度预测问题，通过实证研究，挖掘出结构多样性对流行度预测的指示作用，并在此基础上建立模型进行预测，性能比经典流行度预测方法显著提高。我们的发现为更好地理解社交网络上消息流行度预测乃至信息扩散机制问题提供了一个崭新的视角，具有重要的理论和应用价值。

结语

[1] 一个以科技为主的新闻站点，与一般新闻网站不同的是，在 digg 中用户可以提交新闻并订阅新闻，当订阅数达到一定数量后，digg 算法将自动把新闻加入首页

本文从社会推荐的后验效用、影响力最大化算法、网络结构分析和消息流行度预测四个方面介绍了我们在社会计算方面的一些研究工作。这些研究只是社会计算领域丰富内容的冰山一角。大量社会感知数据汇聚了人类的关系、行为、言论、情感等，是人类社会的数字足迹。人类社会的组织原则、社会规范、活动规律、行为模式都蕴含在社会感知数据中。在当今大数据时代，我们有着前所未有的机遇来开展社会计算。传统的社会学、心理学和认知科学，在大数据背景下有了新驱动力，被注入了新的活力，衍生出了诸如计算社会学、认知计算等交叉学科或领域。社会计算的大幕早已拉开，大数据时代的到来将社会计算推向了高潮，未来的几年内大数据驱动的社会计算将会催生出我们难以预料的成果，人类社会将和自然一样成为自然科学的研究范畴。大数据时代的社会计算，立足数据、面向社会，是大数据的主要应用场景，也必将产生影响深远的研究结果。

参考文献：

[1]. Junming Huang, Xue-Qi Cheng, Hua-Wei Shen, Tao Zhou, Xiaolong Jin. Exploring social influence via posterior effect of word-of-mouth recommendations. Proceedings of the 5th ACM International Conference on Web Search and Data Mining (WSDM), pp. 573-582, February 2012, Seattle, Washington.

[2]. Suqi Cheng, Huawei Shen, Junming Huang, Xueqi Cheng. StaticGreedy: solving the apparent scalability-accuracy dilemma in influence maximization. Submitted to KDD 2013.

[3]. Hua-Wei Shen, Xue-Qi Cheng, Jia-Feng Guo. Exploring the structural regularities in networks. Physical Review E, 84(5):056111, 2011.

[4]. Peng Bao, Hua-Wei Shen, Junming Huang, Xue-Qi Cheng. Popularity prediction in microblogging network: an case study on Sina weibo. Proceedings of the 22nd International Conference on World Wide Web (WWW 2013), May 2013, Rio de Janeiro, Brazil.

[5]. D. Kempe, J. M. Kleinberg, E. Tardos. Maximizing the spread of influence through a social network. In Proceedings of the 9th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2003), pages 137-146, 2003.

[6]. J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, N. S. Glance. Cost-effective outbreak detection in networks. In Proceedings of the 13th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2007), pages 420-429, 2007.

[7]. A. Goyal, W. Lu, and L. V. S. Lakshmanan. CELF++: optimizing the greedy algorithm for influence maximization in social networks. In Proceedings of the 20th International Conference on World Wide Web (WWW 2011), pages 47-48, 2011.

[8]. W. Chen, Y. Wang, S. Yang. Efficient influence maximization in social networks. In Proceedings of the 15th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2009), pages 199-207, 2009.

[9]. W. Chen, Y. Yuan, L. Zhang. Scalable influence maximization in social networks under the linear threshold model. In Proceedings of the 10th IEEE International Conference on Data Mining (ICDM 2010), pages 88-97, 2010.

[10]. M. E. J. Newman, E. A. Leicht. Mixture models and exploratory analysis in networks. Proceedings of National Academy Sciences, 104(23): 9564–9569, 2007.

[11]. Wei Ren, Guiying Yan, Xiaoping Liao, Lan Xiao. Simple probabilistic algorithm for detecting community structure. Physical Review E, 79: 036111, 2009.

[12]. G. Szabo, B. A. Huberman. Predicting the popularity of online content. Communications of the ACM. 53(8): 80-88, 2010.

作者简介：

沈华伟： 中国科学院计算技术研究所、副研究员、网络数据研究中心社会计算基础研究负责人
shenhuawei@ict.ac.cn

程学旗： 中国科学院计算技术研究所、研究员、网络数据研究中心主任、计算所副总工、所长助理